

# Zero-Shot Action Recognition by Word-Vector Embedding

Xun Xu · Timothy Hospedales · Shaogang Gong

Received: date / Accepted: date

**Abstract** The number of categories for action recognition is growing rapidly and it has become increasingly hard to label sufficient training data for learning conventional models for all categories. Instead of collecting ever more data and labelling them exhaustively for all categories, an attractive alternative approach is “zero-shot learning” (ZSL). To that end, in this study we construct a mapping between visual features and a semantic descriptor of each action category, allowing new categories to be recognised in the absence of any visual training data. Existing ZSL studies focus primarily on still images, and attribute-based semantic representations. In this work, we explore word-vectors as the shared semantic space to embed videos and category labels for ZSL action recognition. This is a more challenging problem than existing ZSL of still images and/or attributes, because the mapping between the semantic space and video space-time features of actions is more complex and harder to learn for the purpose of generalising over any cross-category domain shift. To solve this generalisation problem in ZSL action recognition, we investigate a series of synergistic improvements to the standard ZSL pipeline. First, we enhance significantly the semantic space mapping by proposing manifold-regularised regression and data augmentation strategies. Second, we evaluate two existing post processing strategies (transductive self-training and hubness correction), and show that they are complemen-

tary. We evaluate extensively our model on a wide range of human action datasets including HMDB51, UCF101, OlympicSports, CCV and TRECVID MED 13. The results demonstrate that our approach achieves the state-of-the-art zero-shot action recognition performance with a simple and efficient pipeline, and without supervised annotation of attributes. Finally, we present in-depth analysis into why and when zero-shot works, including demonstrating the ability to predict cross-category transferability in advance.

**Keywords** Zero-Shot Action Recognition · Zero-Shot Learning · Semantic Embedding · Semi-Supervised Learning · Transfer Learning · Action Recognition

## 1 Introduction

Action recognition is of established importance in the computer vision community due to its potential applications in video retrieval, surveillance and human machine interaction [Aggarwal and Ryoo \(2011\)](#). However the need for increasing coverage and finer classification of human actions means the number and complexity of action categories of interest for recognition is growing rapidly. For example, action recognition dataset size and number of categories has experienced constant growth since the classic KTH Dataset [Schuldt et al \(2004\)](#) (6 classes, 2004): Weizmann Dataset [Blank et al \(2005\)](#) (9 classes, 2005), Hollywood2 Dataset [Marszalek et al \(2009\)](#) (12 classes, 2009), Olympic Sports Dataset [Niebles \(2010\)](#) (16 classes, 2010), HMDB51 [Kuehne et al \(2011\)](#) (51 classes, 2011) and UCF101 [Soomro et al \(2012\)](#) (101 classes, 2012). The growing number and complexity of actions result in: (1) Enormous human effort is required to collect and label large quantities of video data for

Xun Xu, Timothy Hospedales and Shaogang Gong  
Queen Mary, University of London  
E-mail: xun.xu@qmul.ac.uk

Timothy Hospedales  
E-mail: t.hospedales@qmul.ac.uk

Shaogang Gong  
E-mail: s.gong@qmul.ac.uk

learning. Moreover, compared to image annotation, obtaining each annotated action clip is more costly as it typically requires some level of spatio-temporal segmentation from the annotator. (2) The growing number of categories eventually begins to pose ontological difficulty, about how to structure and define distinct action categories as they grow more fine-grained and inter-related [Jiang et al \(2015\)](#). In this work, we explore methods which do not explicitly create models for new action categories from manually annotated training data, but rather dynamically construct recognition models by combining past experience in language together with knowledge transferred from already labelled existing action categories.

The “zero-shot learning” (ZSL) paradigm [Lampert et al \(2009\)](#); [Fu et al \(2012\)](#); [Socher and Ganjoo \(2013\)](#) addresses this goal by sharing information across categories; and crucially by allowing recognisers for novel/unseen/ testing categories to be constructed based on a semantic *description* of the category, without any labelled *visual* training samples. ZSL methods follow the template of learning a general mapping between a visual feature and semantic descriptor space from known/seen/ training data <sup>1</sup>. In the context of zero-shot action recognition, ‘semantic descriptor’ refers to an action class description that can be specified by a human user, either manually, or with reference to existing knowledge bases, e.g. wikipedia. The ZSL paradigm is most commonly realised by using class-attribute descriptors [Lampert et al \(2014\)](#); [Liu et al \(2011\)](#); [Fu et al \(2015a\)](#) to bridge the semantic gap between low-level features (e.g. MBH or SIFT) and categories. Attributes are mid-level concepts that transcend class boundaries [Lampert et al \(2009\)](#), allowing each category or instance to be represented as a binary [Lampert et al \(2009\)](#); [Liu et al \(2011\)](#) or continuous [Fu et al \(2014a\)](#) vector. Visual attribute classifiers are learned for a set of known categories, and then a human can create recognisers for novel categories by specifying their attributes. With a few exceptions [Liu et al \(2011\)](#); [Fu et al \(2015a\)](#); [Xu et al \(2015\)](#), this paradigm has been applied to images rather than video action recognition.

An emerging alternative to attribute-based ZSL is *unsupervised* semantic embeddings [Socher and Ganjoo \(2013\)](#); [Fu et al \(2014a\)](#); [Habibian et al \(2014b\)](#); [Fu et al \(2015b\)](#); [Xu et al \(2015\)](#); [Akata et al \(2015\)](#). Unsupervised semantic embedding spaces refer to intermediate representations which can be automatically constructed from existing unstructured knowledge-

bases (such as wikipedia text), rather than manually specified attributes. The most common approaches [Socher and Ganjoo \(2013\)](#); [Fu et al \(2014a, 2015b\)](#); [Xu et al \(2015\)](#); [Akata et al \(2015\)](#) are to exploit a distributed vector representation of words produced by a neural network [Mikolov et al \(2013\)](#) trained on a large text corpus in an unsupervised manner. Regressors (cf classifiers in the attribute space), are trained on the known dataset to map low-level visual features into this semantic embedding space. Zero-shot recognition is subsequently performed by mapping novel category visual instances to the embedding space via the regression, and matching these to the vector representation of novel class names (e.g. by nearest neighbour). Several properties make the embedding space approaches preferable to the attribute-based ones: (1) A manually pre-defined attribute ontology is not needed as embedding space is learned in an unsupervised manner. (2) Novel categories can be defined trivially by *naming* them, without the requirement to exhaustively define each class in terms of a list of attributes – which grows non-scale-ably as the breadth of classes to recognise grows [Fu et al \(2014a\)](#); [Akata et al \(2015\)](#). (3) Semantic embedding allows easier exploitation of information sharing across datasets [Xu et al \(2015\)](#); [Habibian et al \(2014b\)](#) because category names from multiple datasets can be easily projected into a common embedding space, while attribute spaces are usually dataset specific, with datasets having incompatible attribute schemas (e.g. UCF101 [Jiang et al \(2013\)](#) and Olympic Sports [Liu et al \(2011\)](#) have disjoint attribute sets).

### The domain shift problem for ZSL of actions

Although embedding-based ZSL is an attractive paradigm, it has rarely previously been demonstrated in zero-shot action recognition. This is in part because of the pervasive challenge of learning mappings, that generalize across the train-test semantic gap [Fu et al \(2015a\)](#); [Romera-Paredes and Torr \(2015\)](#). In ZSL, the train-test gap is more significant than conventional supervised learning because the training and testing classes are *disjoint*, i.e. completely different without any overlap. A serious *domain-shift* [Pan and Yang \(2010\)](#) problem results: mapping from low-level visual feature to semantic embedding trained on a known class data will generalise poorly to novel class data, since the data distributions for the underlying categories are different. This violates the assumptions of supervised learning methods and results in poor performance. The domain shift problem – analysed empirically in [Fu et al \(2015a\)](#); [Dinu et al \(2015\)](#), and theoretically in [Romera-Paredes and Torr \(2015\)](#) – is worse for action than still image recognition because of the greater complexity of cate-

<sup>1</sup> We use *known*, *seen* and *training* interchangeably to refer to the categories with labeled visual training examples and *novel*, *unseen* and *testing* interchangeably to refer to the categories to be recognized without any labeled training samples.

gories in visual space-time features and the mapping of space-time features to semantic embedding space.

**Our Solutions** In this work, we explore four potential solutions to ameliorate the domain shift challenge in ZSL for action recognition as shown in Fig. 1, and achieve better zero-shot action recognition: (1) The first strategy we consider aims to improve the generalisation of the embedding space mapping. We explore **manifold regularisation** (aka semi-supervised learning) to learn a regressor which exploits a regulariser based on the unlabelled data to learn a smoother regressor that better generalises to novel testing classes. Manifold regularisation [Belkin et al \(2006\)](#) is established in semi-supervised learning to improve generalisation of predictions on testing data, but this is more important in ZSL since the gap between training and testing data is even bigger due to disjoint categories. To the best of our knowledge this is the first transductive use of unlabelled data for zero-shot learning at training time. (2) The second strategy we consider is **data augmentation** (aka cross-dataset transfer learning) [Pan and Yang \(2010\)](#); [Shao et al \(2015\)](#). The idea is that by simultaneously learning the regressors for multiple action datasets, a more representative sample of input action data is seen, and thus a more generalizable mapping from the visual feature to the semantic embedding space is learned. This is straightforward to achieve with semantic embedding-based ZSL because the datasets and their category name word-vectors can be directly aggregated. In contrast, it is non-trivial with attribute-based ZSL due to the need to develop a universal attribute ontology for all datasets. Besides these two new considerations to expand the embedding projection, we also evaluate two existing post-processing heuristics to reduce the effect of domain-shift in ZSL. These include (3) **self-training**, which adapts test-class descriptors based on unlabeled testing data to bridge the domain shift [Fu et al \(2014b\)](#) and (4) **Hubness correction** which re-ranks the test-data’s match to novel class descriptions in order to avoid the bias toward ‘hub’ categories induced by domain shift [Dinu et al \(2015\)](#).

By exploring manifold regularization, data augmentation, self-training, and hubness correction, our word-vector embedding approach outperforms consistently conventional zero-shot approaches on all contemporary action datasets (HMDB51, UCF101, Olympic Sports, CCV and USAA). Moreover, on a more relaxed multi-shot setting, our representation is comparable with using low-level features directly. While achieving state-of-the-art performance, the semantic embedding space is constructed in an unsupervised manner, unlike attribute methods which require extensive supervision [Fu et al \(2014a\)](#); [Akata et al \(2015\)](#) (i.e., attribute anno-

tation). Because of a closed-form solution to visual feature to semantic space mapping our method is also very simple to implement and efficient to run on large video databases, unlike many other recent ZSL methods that are significantly more complex and slower [Habibian et al \(2014b\)](#); [Fu et al \(2014a, 2015a\)](#); [Liu et al \(2011\)](#); [Zhao et al \(2013\)](#).

**New Insights** In order to better understand ZSL, for the first time, this study performs a detailed analysis of the relationship between training and testing classes for zero-shot learning, revealing the causal connection between known and novel category recognition performance. This provides a deeper understanding of why and when ZSL works, and sheds light on how to best construct training datasets for effective ZSL in future research and applications.

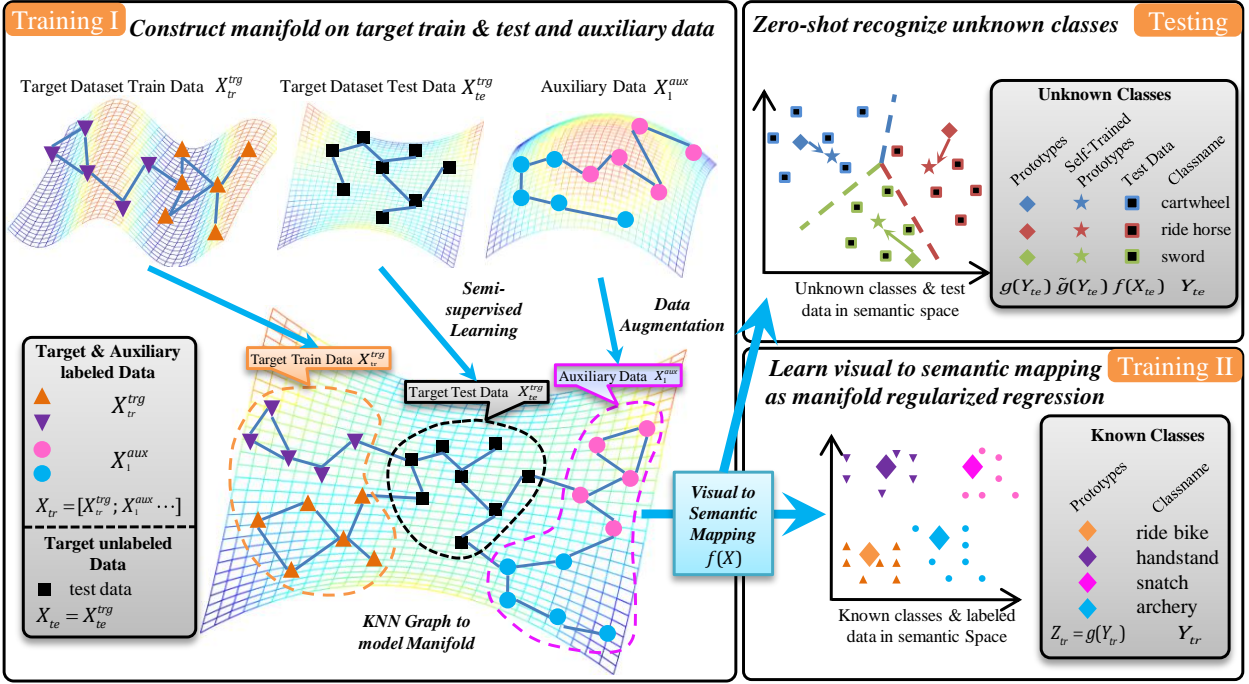
**Contributions** Our key contributions are threefold: (1) We explore jointly four mechanisms for expanding ZSL by addressing its domain-shift challenge, including two existing post processing mechanisms and two new extensions to the embedding mapping itself. Our model is both *closed-form* in solving the visual to semantic mapping and *unsupervised* in constructing the semantic embeddings. (2) We show extensive experiments to demonstrate a very simple implementation of this closed-form model that both runs very quickly and is capable of achieving the state-of-the-art ZSL performance on contemporary action/event datasets. (3) We provide new insight, for the first time, into the underlying factors affecting the efficacy of ZSL.

## 2 Related Work

### 2.1 Action Recognition

Video action recognition is now a vast and established area in computer vision and pattern recognition due to the wide application in video surveillance, interaction between human and electronic devices. Extensive surveys of this area are conducted by [Aggarwal and Ryoo \(2011\)](#); [Poppe \(2010\)](#). Recent progress in this area is attributed to densely tracking points and computing hand-crafted features which are fed into classical supervised classifiers (e.g. SVM) for recognition [Wang et al \(2015\)](#).

**Human Action Datasets** Video datasets for action recognition analysis have experienced constant developing. Early datasets focus on simple and isolated human actions performed by a single person, e.g. KTH [Schuldt et al \(2004\)](#) (2004) and Weizmann [Blank et al \(2005\)](#) (2005) datasets. Due to the growth of internet video sharing, e.g. YouTube and Vimeo, action datasets col-



**Fig. 1** We have labelled data in target dataset  $X_{tr}^{trg}$  and auxiliary dataset  $X_1^{aux}$  and unlabelled data in target dataset  $X_{te}^{trg}$ . The objective is to use all labelled information to help classify unlabelled data into a set of pre-defined categories (aka unknown classes). Specifically, in the training phase I, target labelled data  $X_{tr}^{trg}$  is first augmented by data from auxiliary dataset  $X_1^{aux}$  to form all labelled dataset  $X_{tr}$ . We construct a K Nearest Neighbour (KNN) graph on all labelled and unlabelled data in visual feature space to model the underlying manifold structure. In the training phase II, prototypes for known classes are generated by semantic embedding  $Z_{tr} = g(Y_{tr})$ . Then we learn a visual to semantic mapping  $f : X_{tr} \rightarrow Z_{tr}$  as manifold regularized regression. In the testing phase, prototypes for unknown classes are first generated by semantic embedding  $g(Y_{te})$ . Then target test/unlabelled data  $X_{te}$  are projected into semantic space via  $f(X)$ . Finally simple nearest neighbour (NN) classifier is adopted to categorize test data as the label of closest prototype. On top of NN classifier, self-training and hubness corrections are adopted at testing phase to further improve mitigate domain shift problem. With this framework we achieve the state-of-the-art performance on zero-shot action recognition tasks.

lected from online repositories are emerging, e.g. Olympic Sports [Niebles \(2010\)](#) in 2010, HMDB51 [Kuehne et al \(2011\)](#) in 2011 and UCF101 [Soomro et al \(2012\)](#) in 2012. To recognize more complex events with interactions between people and objects, event datasets including Columbia Consumer Video dataset (CCV) [Jiang et al \(2011\)](#) and the TRECVID Multimedia Event Detection (MED) dataset [Over et al \(2014\)](#) are becoming popular.

**Feature Representation** Local space-time feature approaches have become the prevailing strategies due to not requiring non-trivial object tracking and segmentation. In these approaches, local interest points are first detected [Laptev \(2005\)](#) or densely sampled [Wang et al \(2015\)](#). Visual descriptors invariant to clutter, appearance and scale are calculated in a spatiotemporal volume formed by the interest points. Different visual descriptors have been proposed to capture the texture, shape and motion information, including 3D-SIFT [Sco-vanner et al \(2007\)](#), HOG3D [Klaser et al \(2008\)](#) and local trinary patterns [Yeffe and Wolf \(2009\)](#). Among

these, dense trajectory features with HOG, HOF and MBH descriptors [Wang et al \(2013\)](#) and its variant improved trajectory features [Wang et al \(2015\)](#) produce state-of-the-art performance on action recognition. Therefore, we choose improved trajectory feature (ITF) for our low-level feature representation.

## 2.2 Zero-Shot Learning

Zero-shot learning aims to achieve dynamic construction of classifiers for novel classes at testing time based on semantic descriptors provided by humans or existing knowledge bases, rather than labeled examples. This approach was popularised by the early studies [Larochelle et al \(2008\)](#); [Palatucci et al \(2009\)](#); [Lampert et al \(2009\)](#). Since then numerous studies have been motivated to investigate ZSL due to the scalability barrier of exhaustive annotation for supervised learning, and the desire to emulate the human ability to learn *from description* with few or no examples.



**ZSL Architectures** Various architectures have been proposed for zero-shot recognition of classes  $Y$  given data  $X$ . Sequential architectures Lampert et al (2009); Fu et al (2014a, 2015a); Liu et al (2011); Zhao et al (2013); Lazaridou et al (2014) setup classifier/regressor mappings  $Z = f(X)$  to predict semantic representations  $Z$ , followed by a recognition function  $Y = r(Z)$ . Visual feature mapping  $f(\cdot)$  is learned from training data and assumed to generalise, and the recogniser is given by the human or external knowledge. Converging architectures Akata et al (2015); Yang and Hospedales (2015); Romera-Paredes and Torr (2015) setup energy functions  $E(X, Z)$  which are positive when  $X$  and  $Z$  are from matching classes and negative otherwise. In this work, we adopt a sequential regression approach for simplicity and efficiency.

**Attribute Embeddings** The most popular intermediate representation for ZSL has been attributes, where categories are specified in terms of a vector of binary Lampert et al (2009); Liu et al (2011); Zhao et al (2013) or continuous Fu et al (2014a); Akata et al (2015); Romera-Paredes and Torr (2015) attributes. However, this approach suffers inherently from the need to agree upon a universal attribute ontology, and the scalability barrier of manually defining each new class in terms of an attribute ontology that grows with breadth of classes considered Fu et al (2014a); Akata et al (2015).

**Word-Vector Embeddings** While other representations including taxonomic Akata et al (2015), co-occurrence Gan et al (2015); Mensink et al (2014); Habibian et al (2014b) and template-based Larochelle et al (2008) have been considered, word-vector space ZSL Fu et al (2015a); Akata et al (2015); Xu et al (2015); Lazaridou et al (2014) has emerged as the most effective unsupervised alternative to attributes. In this approach, the semantic class descriptor  $Z$  is generated automatically from existing unstructured text knowledge bases such as wikipedia. In practice, this often means the target  $Z$  of mapping  $Z = f(X)$  is given by the internal representation of a text modelling neural network Mikolov et al (2013). This can be more intuitively understood as encoding each class name in terms of a vector describing its co-occurrence frequency with other terms in a text corpus Lazaridou et al (2014).

**Domain-Shift** Every ZSL method suffers from the issue of domain shift between the training class on which the mapping  $f(\cdot)$  or energy function  $E(\cdot, \cdot)$  is trained, and the disjoint set of testing classes to which it is tested on. Although this is a major reason why it is hard to obtain competitive results with ZSL strategies, it is only recently this problem has been studied explicitly Dinu et al (2015); Fu et al (2015a); Romera-Paredes

and Torr (2015). In this work, we focus primarily on how to mitigate this domain-shift problem in ZSL for action recognition. That is, by making the training data more representative thus learning a more general visual feature to semantic space mapping (dataset augmentation), transductively exploiting both labelled and unlabelled data manifold to learn an embedding mapping that generalises better to the testing data (manifold regularized regression), and post-processing corrections to adapt (self-training) the classifier at the testing time therefore to improve its robustness (hubness correction) to domain shift. While transductive Dinu et al (2015); Fu et al (2015a); Xu et al (2015) strategies have been exploited before as post-processing, this is the first time it have been exploited for learning the embedding itself via manifold regression.

### 2.3 ZSL for Action Recognition

Despite clear appeal from ZSL, few studies have considered it for action recognition. Early attribute-centric studies took latent SVM Liu et al (2011) and topic model Fu et al (2014a); Zhao et al (2013) approaches, neither of which are very scalable for large video datasets. Thus more recent studies have started to consider unsupervised embeddings including semantic relatedness Gan et al (2015) and word-vectors Xu et al (2015). However, most prior ZSL action recognition studies do not evaluate against a wide range of realistic set of contemporary action recognition benchmarks, restricting themselves to a single dataset of USAA Fu et al (2014a); Zhao et al (2013), or Olympic Sports Liu et al (2011). In this work, we fully explore word-vector-based zero-shot action recognition, and demonstrate its superiority to attribute-based approaches, despite the latter’s supervised ontology construction.

## 3 Methodology

To formalise the problem a list of notations are first given in Table 1. We have a training video set  $T_{tr} = \{X_{tr}, Y_{tr}\}$  where  $X_{tr} = \{x_i\}_{i=1 \dots n_l}$  is the set of  $d_x$  dimensional low-level features, e.g. Fisher Vector encoded MBH and HOG. For each of the  $n_l$  labelled training videos  $Y_{tr} = \{y_i\}_{i=1 \dots n_l}$  is the class names/labels of each instance, e.g. “brush hair” and “handwalk”. We also have a set of testing videos  $T_{te} = \{X_{te}, Y_{te}\}$  with  $n_u$  unlabelled testing video instances. The goal of ZSL is to learn to recognise videos in  $X_{te}$  whose classes  $Y_{te}$  are disjoint from any seen data at training time:  $Y_{tr} \cap Y_{te} = \emptyset$ .

**Table 1** Basic notations.

Notation	Description
$X \in \mathbb{R}^{d_x \times N}; x_i$	Visual feature matrix for N instances; Column representing the $i$ -th instance
$Y \in \mathbb{Z}^{1 \times N}; y_i$	Integer class labels for N instances; Column representing the $i$ -th instance
$Z \in \mathbb{R}^{d_z \times N}; z_i$	Semantic embedding for N instances; Column representing the $i$ -th instance
$K \in \mathbb{R}^{N \times N}$	Kernel matrix
$A \in \mathbb{R}^{d_z \times N}$	Regression coefficient matrix
$f : X \rightarrow Z$	Visual to semantic mapping function
$g : Y \rightarrow Z$	Class name embedding function
$\lambda_A \in \mathbb{R}$	Ridge regression regularizer
$\lambda_I \in \mathbb{R}$	Manifold regression regularizer
$N_K^G \in \mathbb{Z}^+$	KNN Graph parameter for manifold regularizer
$N_K^S \in \mathbb{Z}^+$	KNN parameter for Self-Training procedure

### 3.1 Semantic Embedding Space

To bridge the gap between disjoint training and testing classes, we establish a semantic embedding space  $Z$  based on word-vectors. In particular we use a neural network Mikolov et al (2013) trained on a 100 billion word corpus to realise a mapping  $g : Y \rightarrow Z$  that produces a unique  $d_z$  dimensional encoding vector of each dictionary word.

**Compound Names** The above procedure only deals with class names that are unigram dictionary words. To process compound names commonly occurring in action datasets, e.g. “brush hair” or “ride horse”, that do not exist as individual tokens in the corpus, we exploit compositionally of the semantic space Mitchell and Lapata (2008). Various composition methods have been proposed Mitchell and Lapata (2008); Milajevs et al (2014) including additive, multiplicative and others, but our experiments showed no significant to using others besides addition, so we stick with simple additive composition.

Suppose the  $i$ th class name  $y_i$  is composed of words  $\{y_{ij}\}_{j=1 \dots w}$ . We generate a single  $d_z$  dimensional vector  $z$  out of the word-vector  $y_i$  by summing the word-vectors for constituent words  $\{y_{ij}\}$ :

$$z_i = \frac{1}{N} \cdot \sum_{j=1}^N g(y_{ij}) \quad (1)$$

### 3.2 Visual to Semantic Mapping

**Mapping by Regression:** In order to map video features into the semantic embedding space constructed above, we train a regression model  $f : X \rightarrow Z$  from  $d_x$  dimensional low-level visual feature space to the  $d_z$  dimensional embedding space. The regression is trained using training instances  $X_{tr} = \{x_i\}_{i=1 \dots n_l}$  and the corresponding embedding  $Z_{tr} = g(Y_{tr})$  of the instance class name  $y$  as the target value. Various methods have

previously been used for this task including linear support vector regression (SVR) Fu et al (2014a, 2015a); Xu et al (2015) and more complex multi-layer neural networks Socher and Ganjoo (2013); Lazaridou et al (2014); Yang and Hospedales (2015). Since we will use fisher vector encoding Perronnin et al (2010) for features  $X$ , we can easily apply simple linear regression for  $f(\cdot)$ . Specifically, we use  $l_2$  regularised linear regression (aka ridge regression) to learn the visual to semantic mapping.

**Kernel Ridge Regression:** The fisher vector encoding generates a very high dimensional feature  $D = 2 \times d_{desc} \times k$  where  $k$  is number of components in the Gaussian Mixture Model (GMM) and  $d_{desc}$  is the dimension of raw descriptors. This usually results in many more feature dimensions than training samples. Thus we use the representer theorem Scholkopf and Smola (2002) and formulate a kernelized ridge regression with a linear kernel defined by Eq. (2):

$$K(x_i, x_j) = \sum_{d=1}^{d_x} (x_{id} \cdot x_{jd}) \quad (2)$$

With the classical representer theorem, visual feature  $x$  can be projected into semantic space via Eq. (3) where  $A_j$  is the  $j$ th column of regression parameter matrix  $A$ .

$$f(x) = \sum_{j=1}^{n_l} A_j K(x, x_j) \quad (3)$$

To improve the generalisation of the resulting regressor, we add the  $l_2$  regulariser  $\|f\|_K^2 = Tr(AKA^T)$  to reduce overfitting by penalising extreme values in the regression matrix. This gives the kernel ridge regression loss:

$$\begin{aligned} & \min_f \frac{1}{n_l} \sum_{i=1}^{n_l} \|z_i - f(x_i)\|_2^2 + \gamma \|f\|_K^2 \\ & \min_f \frac{1}{n_l} \sum_{i=1}^{n_l} (z_i - f(x_i))^T (z_i - f(x_i)) + \gamma Tr(AKA^T) \\ & \min_A \frac{1}{n_l} Tr((Z - AK)^T (Z - AK)) + \gamma Tr(AKA^T) \end{aligned} \quad (4)$$

where the regression targets are generated by the vector representation of each class name  $z_i = g(y_i)$  and  $Z = [z_1 \ z_2 \ \dots]_{d_z \times n_l}$ ,  $A$  is the  $d_z \times n_l$  regression coefficient matrix and  $K$  is the  $n_l \times n_l$  kernel matrix. The loss function is convex with respect to the  $A$ . Taking derivatives w.r.t  $A$  and setting the gradient to 0 leads to the

following closed-form solution where  $\mathbf{I}$  is the identity matrix.

$$A = Z(K + \gamma_A n_l \mathbf{I})^{-1} \quad (5)$$

The above mapping by Kernel Ridge Regression provides a simple solution to embed visual instances into semantic space. However the simple ridge regression only considers limited labelled training data  $X_{tr}$  without exploiting the underline structure of the manifold on both labelled and unlabelled data nor any additional related labelled data from other datasets. In the following sections, we introduce two approaches to improve the quality of mapping: (1) *Manifold-Regularized Regression* and (2) *Data Augmentation*.

### 3.2.1 Manifold-Regularized Regression

As discussed earlier, conventional regularisation provides poor ZSL due to disjoint training and testing classes. To improve recognition of testing classes, we explore transductive semi-supervised regression. The idea is to exploit unlabelled testing data  $X_{te}$  to discover the manifold structure in the zero-shot classes, and preserve this structure in the semantic space after visual-semantic mapping. Therefore, this is also known as manifold regularization.

To that end, we introduce manifold laplacian regularization [Belkin et al \(2006\)](#) into the ridge regression formulation. This additional regularization term ensures that if two videos are close to each other in the visual feature space, this relationship should be kept in the semantic space as well.

We model the manifold by constructing a symmetric  $K$  nearest neighbour graph  $W$  on the all  $n_l + n_u$  instances where  $n_l = |T_{tr}|$  denotes the number of labelled training instances and  $n_u = |T_{te}|$  denotes the number of unlabelled testing instances. Let  $D$  be a diagonal matrix with  $D_{ii} = \sum_{j=1}^{n_l+n_u} W_{ij}$ , we get the graph laplacian matrix  $L = D - W$ . The manifold regularizer is then written as:

$$\begin{aligned} \|f\|_I^2 &= \frac{1}{2} \sum_{i,j}^{n_l+n_u} W_{ij} \|f(x_i) - f(x_j)\|_2^2 \\ &= \frac{1}{2} \sum_{i,j} W_{ij} f^T(x_i) f(x_i) + \frac{1}{2} \sum_{i,j} W_{ij} f^T(x_j) f(x_j) \\ &\quad - \sum_{i,j} W_{ij} f^T(x_i) f(x_j) \\ &= \sum_i D_{ii} f^T(x_i) f(x_i) - \sum_{i,j} W_{ij} f^T(x_i) f(x_j) \end{aligned} \quad (6)$$

Further denoting  $\mathbf{f} = [f(x_1) \ f(x_2) \cdots \ f(x_{n_l+n_u})] = AK$ . Eq. (6) can be rewritten as:

$$\begin{aligned} \|f\|_I^2 &= Tr(\mathbf{f}^T \mathbf{f} D) - Tr(\mathbf{f}^T \mathbf{f} W) \\ &= Tr(\mathbf{f}^T \mathbf{f} L) \\ &= Tr(K^T A^T AKL) \end{aligned} \quad (7)$$

where  $K$  is a  $(n_l + n_u) \times (n_l \times n_u)$  dimensional kernel matrix constructed upon all labelled and unlabelled instances via Eq (2). Combining all regularization terms we obtain the overall loss function in Eq (8), where for simplicity we denote  $J = \begin{bmatrix} \mathbf{I}_{n_l \times n_l} & \mathbf{0}_{n_l \times n_u} \\ \mathbf{0}_{n_u \times n_l} & \mathbf{0}_{n_u \times n_u} \end{bmatrix}$  and  $\tilde{Z} = [Z_{tr} \ \mathbf{0}_{d_z \times n_u}]$ . The final loss function can be thus written as:

$$\begin{aligned} \min_f \frac{1}{n_l} \sum_{i=1}^{n_l} \|z_i - f(x_i)\|_2^2 &+ \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(n_l + n_u)^2} \|f\|_I^2 \\ \min_A \frac{1}{n_l} Tr((\tilde{Z} - AKJ)^T (\tilde{Z} - AKJ)) &+ \gamma_A Tr(AKA^T) \\ &+ \frac{\gamma_I}{(n_l + n_u)^2} Tr(K^T A^T AKL) \end{aligned} \quad (8)$$

The loss function is convex w.r.t. the  $d_z \times (n_l + n_u)$  regression coefficient matrix  $A$ . A closed-form solution to  $A$  can be obtained in the same way as Kernel Ridge Regression.

$$A = \tilde{Z} \left( KJ + \gamma_A n_l \mathbf{I} + \frac{\gamma_I n_l}{(n_l + n_u)^2} KL \right)^{-1} \quad (9)$$

Eq (9) provides an efficient way to learn the visual to semantic mapping due to the closed-form solution compared to alternative iterative approaches [Fu et al \(2014a\)](#); [Habibian et al \(2014b\)](#). At testing time, the mapping can be efficiently applied to project new videos into the embedding with Eq. (3). Note when  $\gamma_I = 0$  manifold regression becomes exactly kernel regression.

### 3.2.2 Improving the Embedding with Data Augmentation

As discussed, the mapping often generalises poorly because: (i) actions are visually complex and ambiguous, and (ii) even a mapping well learned for training categories may not generalise well to testing categories as required by ZSL, because the volume of training data is small compared to the complexity of a general visual to semantic space mapping. The manifold regression described previously ameliorates the latter issues, but we next discuss a complementary strategy of data augmentation (cross-dataset transfer).

Another way to further mitigate both of these problems is by augmentation with any available auxiliary dataset which need not contain classes in common with

the target dataset  $T^{trg}$  in which zero-shot recognition is performed. This will provide more data to learn a better generalising regressor  $z = f(x)$ . We formalize the data augmentation problem as follows. We denote the target dataset as  $T^{trg} = \{X^{trg}, Y^{trg}\}$  split into training set  $T_{tr}^{trg} = \{X_{tr}^{trg}, Y_{tr}^{trg}\}$  and zero-shot testing set  $T_{te}^{trg} = \{X_{te}^{trg}, Y_{te}^{trg}\}$ . Zero-shot recognition is performed on the testing set of target dataset (e.g. HMDB51). There are  $n_{aux}$  other available auxiliary datasets  $T_{i=1 \dots n_{aux}}^{aux} = \{X_i^{aux}, Y_i^{aux}\}$  (e.g. UCF101, Olympic Sports and CCV). We propose a simple but effective approach to improve the regression by merging the target dataset training data and all auxiliary sets. The auxiliary dataset class names  $Y_i^{aux}$  are projected into the embedding space with  $Z_i^{aux} = g(Y_i^{aux})$ . The auxiliary instances  $X_i^{aux}$  are aggregated with the target training data  $X_{tr} = [X_{tr}^{trg} \ X_1^{aux} \ \dots \ X_{n_{aux}}^{aux}]$ ,  $Z_{tr} = [Z_{tr}^{trg} \ Z_1^{aux} \ \dots \ Z_{n_{aux}}^{aux}]$  where  $Z_{tr}^{trg} = g(Y_{tr}^{trg})$  and together used to train the regressor  $f$ .

To write loss function in matrix calculation we denote  $n_l^{trg} = |T_{tr}^{trg}|$ ,  $n_u^{trg} = |T_{te}^{trg}|$ ,  $n_l^{aux} = \sum_i |T_i^{aux}|$ ,

$\tilde{Z}_{tr} = [Z_{tr} \ \mathbf{0}_{d_z \times n_u}]$ ,  $\tilde{K}$  as the  $(n_l^{trg} + n_u^{trg} + n_l^{aux}) \times (n_l^{trg} + n_u^{trg} + n_l^{aux})$  dimensional kernel matrix upon all target and auxiliary data,

$\tilde{J} = \begin{bmatrix} \mathbf{I}_{(n_l^{trg} + n_l^{aux}) \times (n_l^{trg} + n_l^{aux})} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_u^{trg} \times n_u^{trg}} \end{bmatrix}$  and  $\tilde{L}$  as the graph laplacian matrix defined on all target and auxiliary data. The loss function of manifold regularized regression with data augmentation is thus written as:

$$\min_f \frac{1}{(n_{tr}^{trg} + n_l^{aux})} \left( \sum_{i=1}^{n_l^{trg}} \|z_{tr_i}^{trg} - f(x_{tr_i}^{trg})\|_2^2 + \sum_{i=1}^{n_{aux}} \sum_{j=1}^{|T_i^{aux}|} \|z_{ij}^{aux} - f(x_{ij}^{aux})\|_2^2 \right) + \gamma_A \|f\|_K^2 + \frac{\gamma_I}{(n_{tr}^{trg} + n_{te}^{trg} + n_l^{aux})^2} \|f\|_I^2 \quad (10)$$

$$\min_A \frac{1}{(n_{tr}^{trg} + n_l^{aux})} Tr \left( (\tilde{Z}_{tr} - A\tilde{K}\tilde{J})^T (\tilde{Z}_{tr} - A\tilde{K}\tilde{J}) \right) + \gamma_A Tr(A\tilde{K}A^T) + \frac{\gamma_I}{(n_{tr}^{trg} + n_{te}^{trg} + n_l^{aux})^2} Tr(\tilde{K}^T A^T A \tilde{K} \tilde{L})$$

In the same way as before, we obtain the closed-form solution to  $A$ :

$$A = \tilde{Z}_{tr} \left( \tilde{K}\tilde{J} + \gamma_A (n_{tr}^{trg} + n_l^{aux}) \mathbf{I} + \frac{\gamma_I (n_{tr}^{trg} + n_l^{aux})}{(n_{tr}^{trg} + n_{te}^{trg} + n_l^{aux})^2} \tilde{K}\tilde{L} \right)^{-1} \quad (11)$$

Note that by setting  $\gamma_I = 0$  we obtain a kernel ridge regression with only data augmentation.

### 3.3 Zero-Shot Recognition

Given the trained mappings  $f(\cdot)$  and  $g(\cdot)$  we can now complete the zero-shot learning task. To classify a testing instance  $x^* \in X_{te}$ , we apply nearest neighbour matching of the projected testing instance  $f(x^*)$  against the vector representations of all the testing classes  $g(y)$  (named the prototype throughout this paper):

$$\hat{y} = \arg \min_{y \in Y_{te}} \|f(x^*) - g(y)\| \quad (12)$$

Distances in such embedding spaces have been shown to be best measured using the cosine metric [Mikolov et al \(2013\)](#); [Fu et al \(2014a\)](#). Thus we  $l_2$  normalise each data point, making euclidean distance effectively equivalent to cosine distance in this space.

#### 3.3.1 Ameliorating Domain Shift by Post Processing

In the previous two sections we introduced two methods to improve the embedding  $f$  for ZSL. In this section we now discuss two post-processing strategies to further reduce the impact of domain shift.

**Self-training for Domain Adaptation** The domain shift induced by applying  $f(\cdot)$  trained on  $X_{tr}$  to data of different statistics  $X_{te}$  means the projected data points  $f(X_{te})$  do not lie neatly around the corresponding class projections/prototypes  $g(Y_{te})$  [Fu et al \(2015a\)](#). To ameliorate this domain shift, we explore transductive self-training to adjust unseen class prototypes to be more comparable to the projected data points. For each category prototype  $g(y^*)$ ,  $y^* \in Y_{te}$  we search for the  $N_K^{st}$  nearest neighbours among the unlabelled testing instance projections, and re-define the adapted prototype  $\tilde{g}(y^*)$  as the average of those  $N_K^{st}$  neighbours. Thus if  $NN_K(g(y^*))$  denotes the set of  $K$  nearest neighbours of  $g(y^*)$ , we have:

$$\tilde{g}(y^*) := \frac{1}{N_K^{st}} \sum_{f(x^*) \in NN_K(g(y^*))} f(x^*) \quad (13)$$

The adapted prototypes  $\tilde{g}(y^*)$  are now more directly comparable with the testing data for matching using Eq. (12).

**Hubness Correction** One practical effect of the ZSL domain shift was elucidated in [Dimu et al \(2015\)](#), and denoted the ‘Hubness’ problem. Specifically, after the domain shift, there are a small set of ‘hub’ test-class prototypes that become nearest or  $K$  nearest neighbours to the majority of testing samples in the semantic space, while others are NNs of no testing instances. This results in poor accuracy and highly biased predictions with the majority of testing examples being assigned



to a small minority of classes. We therefore explore the simple solutions proposed by Dinu et al (2015) which takes into account the global distribution of zero-shot samples and prototypes. This method is transductive as with self-training and manifold-regression. Specifically, we considered two alternative approaches: *Normalized Nearest Neighbour* (NRM) and *Globally Corrected* (GC).

The NRM approach eliminates the bias towards hub prototypes by normalizing the distance of each prototype to all testing samples prior to performing Nearest Neighbour classification as defined in Eq (12). More specifically, denote the distance between prototype  $y_j$  and testing sample  $\{x_i^*\}_{i=1\dots n_u}$  as  $d_{ij} = \|f(x_i^*) - g(y_j)\|$ . We then  $l_2$  normalize the distances between prototype  $y_j$  and all  $n_u$  testing samples in Eq (14). This normalized distance  $\tilde{d}_{ij}$  replaces the original distance  $d_{ij}$  for doing nearest neighbour matching in Eq. (12).

$$\tilde{d}_{ij} = d_{ij} / \sqrt{\sum_i^{n_u} d_{ij}^2} \quad (14)$$

The alternatively GC approach damps the effect of hub prototypes by using ranks rather than the original distance measures. We denote the function  $Rank(y, x_i^*; x_i^* \in X_{te})$  as the rank of testing sample  $x_i^*$  w.r.t the distance to  $y$ . The rank function always return an integer value between 1 and  $|X_{te}|$ . Thus the label of testing sample  $x_i^*$  can be predicted by Eq (15) in contrast to simple nearest by neighbour Eq (12).

$$\hat{y} = \arg \min_{y \in Y_{te}} Rank(y, x_i^*) \quad (15)$$

### 3.4 Multi-Shot Learning

Although our focus is zero-shot learning, we also note that the semantic embedding space provides an alternative representation for conventional supervised learning. For multi-shot learning, we map all data instances  $X$  into the semantic space using projection  $Z = f(X)$ , and then simply train SVM classifiers with linear kernel using the  $l_2$  normalised projections  $f(X)$  as data. In the testing phase, testing samples are projected into embedding space via the mapping  $f(X)$  and categorised using the SVM classifiers.

## 4 Experiments

### 4.1 Datasets and Settings

**Datasets:** Experiments are performed on 4 popular contemporary action recognition and event detection

datasets including A Large Human Motion Database (HMDB51) Kuehne et al (2011), UCF101 Soomro et al (2012), Olympic Sports Niebles (2010) and Columbia Consumer Video (CCV) Jiang et al (2011). HMDB51 is specifically created for human action recognition. It has 6766 videos from various sources with 51 categories of actions. UCF101 is an action recognition dataset of 13320 realistic action videos, collected from YouTube, with 101 action categories. Olympic Sports is collected from YouTube, and is mainly focused on sports events. It has 783 videos with 16 categories of events. CCV contains 9682 YouTube videos over 20 semantic categories. We illustrate some example frames in Fig. 2. The action/event category names are presented in Table 2. We also evaluate USAA Fu et al (2014a) – a subset of CCV specifically annotated with attributes – in order to facilitate comparison against attribute centric ZSL approaches. In addition to above action/event datasets, we also studied a large complex event dataset - TRECVID MED 2013. There are five components to the dataset including Event Kit training, Background training, test set MED, test set Kindred and Research Set. We use standard test set MED for zero-shot testing data and Event Kit as training data.

**Visual Feature Encoding:** For each video we extract improved trajectory feature (ITF) descriptors Wang and Schmid (2013) and encode them with Fisher Vectors (FV). We first compute ITF with 3 descriptors (HOG, HOF and MBH). We apply PCA to reduce the dimension of descriptors by half which results in descriptors with 198 dimensions in total. Then we randomly sample 256,000 descriptors from each of the 4 action/event dataset and learn a Gaussian Mixture Model with 128 components from the combined training descriptors. Finally the dimension of FV encoded feature is equal to  $d_x = 2 \times 128 \times 198 = 50688$ . The visual feature for TRECVID MED 2013 dataset was extracted using ITF with HOG and MBH descriptors encoded with Fisher Vectors. We use the FV encoded feature provided by Habibian et al (2014b).

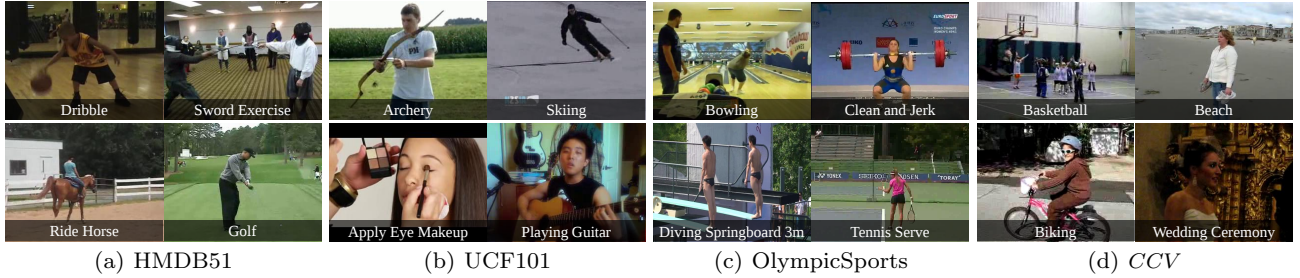
**Semantic Embedding Space:** We adopted the skip-gram neural network model Mikolov et al (2013) trained on the Google News dataset (about 100 billion words). This neural network can then encode any of approximately 3 million unique words as a  $d_z = 300$  dimension vector.

### 4.2 Zero-shot Learning on Actions and Events

**Data Split:** Because there is no existing zero-shot learning evaluation protocol for most existing action

**Table 2** Category names of each dataset.

Dataset	Categories
HMDB51	brush_hair, cartwheel, catch, chew, clap, climb, climb_stairs, dive, draw_sword, dribble, drink, eat, fall_floor, fencing, flic_flac, golf, handstand, hit, hug, jump, kick, kick_ball, kiss, laugh, pick, pour, pullup, punch, push, pushup, ride_bike, ride_horse, run, shake_hands, shoot_ball, shoot_bow, shoot_gun, sit, situp, smile, smoke, somersault, stand, swing_baseball, sword, sword_exercise, talk, throw, turn, walk, wave
UCF101	Apply Eye Makeup, Apply Lipstick, Archery, Baby Crawling, Balance Beam, Band Marching, Baseball Pitch, Basketball Shooting, Basketball Dunk, Bench Press, Biking, Billiards Shot, Blow Dry Hair, Blowing Candles, Body Weight Squats, Bowling, Boxing Punching Bag, Boxing Speed Bag, Breaststroke, Brushing Teeth, Clean and Jerk, Cliff Diving, Cricket Bowling, Cricket Shot, Cutting In Kitchen, Diving, Drumming, Fencing, Field Hockey Penalty, Floor Gymnastics, Frisbee Catch, Front Crawl, Golf Swing, Haircut, Hammer Throw, Hammering, Handstand Pushups, Handstand Walking, Head Massage, High Jump, Horse Race, Horse Riding, Javelin Throw, Juggling Balls, Jump Rope, Jumping Jack, Kayaking, Knitting, Long Jump, Lunges, Military Parade, Mixing Batter, Mopping Floor, Nun chucks, Parallel Bars, Pizza Tossing, Playing Guitar, Playing Piano, Playing Tabla, Playing Violin, Playing Cello, Playing Daf, Playing Dhol, Playing Flute, Playing Sitar, Pole Vault, Pommel Horse, Pull Ups, Punch, Push Ups, Rafting, Rock Climbing Indoor, Rope Climbing, Rowing, Salsa Spins, Shaving Beard, Shotput, Skate Boarding, Skiing, Skijet, Sky Diving, Soccer Juggling, Soccer Penalty, Still Rings, Sumo Wrestling, Surfing, Swing, Table Tennis Shot, Tai Chi, Tennis Swing, Throw Discus, Trampoline Jumping, Typing, Uneven Bars, Volleyball Spiking, Walking with a dog, Wall Pushups, Writing On Board, Yo Yo
Olympic Sports	basketball layup, bowling, clean and jerk, discus throw, hammer throw, high jump, javelin throw, long jump, diving platform 10m, pole vault, shot put, snatch, diving springboard 3m, tennis serve, triple jump, vault
CCV	Basketball, Baseball, Soccer, IceSkating, Skiing, Swimming, Biking, Cat, Dog, Bird, Graduation, Birthday, WeddingReception, WeddingCeremony, WeddingDance, MusicPerformance, NonmusicPerformance, Parade, Beach, Playground

**Fig. 2** Example frames for different action datasets.

and event datasets we propose our own splits<sup>2</sup>. We first propose a 50/50 category split for all datasets. Visual to semantic space mappings are trained on the 50% training categories, and the other 50% are held out unseen for testing time. We randomly generate 50 independent splits and take the mean accuracy and standard deviation for evaluation. Among the 50 splits, all categories are evaluated as testing classes, and the frequency is evenly distributed.

**Evaluation of Components:** To evaluate the efficacy of each component we considered an extensive combination of blocks including manifold regularizer, self-training, hubness correction and data augmentation. Specifically we evaluated the following options for each component.

- **Data Augmentation:** Using only within target dataset training data (X) to learn the embedding  $f(x)$ , or also borrowing data from the auxiliary datasets ( $\checkmark$ ). (Section 3.2.2). For each of the four datasets HMDB51, UCF101, Olympic Sports and CCV, the other three datasets are treated as the auxiliary sets.

- **Embedding:** We compare ridge regression (RR) with manifold regularized ridge regression (MR) (Section 3.2).
- **Self Training:** With ( $\checkmark$ ) or without (X) self-training before matching (Section 3.3.1).
- **Matching Strategy:** We compare conventional NN matching (NN) Eq. (12) versus Normalised Nearest Neighbour (NRM) Eq. (14) and Globally Corrected (GC) matching Eq. (15) (Section 3.3.1).

Based on this breakdown of components, we note that the condition (X-RR-X-NN) is roughly equivalent to the methods in Socher and Ganjoo (2013) and Lazari-dou et al (2014), and the conditions (X-RR-X-GC, X-RR-X-NRM) are roughly equivalent to Dinu et al (2015).

**Metrics:** HMDB, UCF and USAA are classification benchmarks, so we report average accuracy metric. Olympic-Sports and CCV are detection benchmarks, so we report mean average precision (mAP) metrics. NRM and GC do not change the performance of NN for the detection benchmarks, so we do not report these.

**Comparison With Others:** In addition to the above variants of our framework, we also evaluate the following prior state-of-the-art approaches to ZSL. Note that these are based on a supervised embedding using man-

<sup>2</sup> The data split will be released on our website

ually labeled attributes, and as such are only evaluated on UCF, Olympic Sports and CCV where attributes are available.

1. **Direct Attribute Prediction (DAP)** We implement the method of [Lampert et al \(2014\)](#), but using the same FV encoded visual features and linear kernel SVM attribute classifiers  $p(a|x)$ . Recognition is then performed based on attribute posteriors and manually specified attribute descriptor  $p(a|y)$ .
2. **Indirect Attribute Prediction (IAP)** [Lampert et al \(2014\)](#). This differs from DAP by learning a per-category classifier  $p(y|x)$  from training data first and then use the training category attribute-prototype dependency  $p(a|y)$  to obtain attribute estimator  $p(a|x)$ .
3. **Human Actions by Attributes (HAA)** [Liu et al \(2011\)](#). We reproduce a simplified version of this model which exploits the manually labelled attributes  $\{a_m\}$  for zeroshot learning. Similar to DAP, a binary SVM classifier is trained per attribute. In the testing phase, each testing sample is projected into attribute space and then assigned to the closest testing/unknown class based on cosine distance to the class prototype (NN).
4. **Multi-Modal Latent Attribute Topic Model (M2LATM)** [Fu et al \(2014a\)](#). This model fuses multiple features – static (SIFT), motion (STIP) and audio (MFCC) – and exploits both user-defined attributes and discovered latent attributes to facilitate zero-shot learning. We report the results on USAA from [Fu et al \(2014a\)](#).

**Experimental Results:** The full results are presented in Table 3, from which we draw the following conclusions: (i) The simplest approach of directly mapping features to the embedding space (X-RR-X-NN, [Socher and Ganjoo \(2013\)](#); [Lazaridou et al \(2014\)](#)) works reasonably well, suggesting that semantic space is effective as a representation and supports ZSL. (ii) While this simple approach is not clearly better than attribute based approaches [Lampert et al \(2014\)](#); [Liu et al \(2011\)](#), it does not require the latter’s extensive and costly attribute annotation. (iii) Manifold regularisation reliably improves performance compared to conventional ridge regression by reducing the domain shift through considering the unlabeled testing data. (iv) Data augmentation also significantly improves the results by providing a more representative sample of training data for learning the embedding. (v) In line with previous work self-training [Fu et al \(2015a\)](#) and Hubness [Dinu et al \(2015\)](#) post-processing improve results at testing time, and this is complementary with our proposed manifold regular-

isation and data augmentation. (v) Overall the combination of all components (data-augmentation, manifold regularisation, self-training and normalised matching) provides the best performance by reducing the effect of the domain shift. Depending on the dataset, this is comparable or significantly better than the attribute-centric methods, despite the latter’s use of a more costly supervised embedding.

#### 4.3 Zero-shot Learning of Complex Events

In this section, we experiment on the more challenging complex event dataset - TRECVID MED 2013.

**Data Split:** We study the 30 classes of the MED test set, holding out the 20 events specified by the 2013 evaluation scheme for zero-shot recognition, and training on the other 10. We train on the 160 examples per class (on average) in Event Kit Train and test on the 27K examples in MED test, of which only about 1448 videos are the 20 events to be detected<sup>3</sup> This is a challenging setting because the training classes are few in both number and examples per category.

Table 4 Events for training visual to semantic regression

ID	Event Name	ID	Event Name
E001	Attempting a board trick	E002	Feeding an animal
E003	Landing a fish	E004	Wedding ceremony
E005	Working on a woodworking project	E016	Doing homework or studying
E017	Hide and seek	E018	Hiking
E019	Installing flooring	E020	Writing

Table 5 Events for testing zeroshot event detection

ID	Event Name	ID	Event Name
E006	Birthday party	E007	Changing a vehicle tire
E008	Flash mob gathering	E009	Getting a vehicle unstuck
E010	Grooming an animal	E011	Making a sandwich
E012	Parade	E013	Parkour
E014	Repairing an appliance	E015	Working on a sewing project
E021	Attempting a bike trick	E022	Cleaning an appliance
E023	Dog show	E024	Giving directions to a location
E025	Marriage proposal	E026	Renovating a home
E027	Rock climbing	E028	Town hall meeting
E029	Winning a race without a vehicle	E030	Working on a metal crafts project

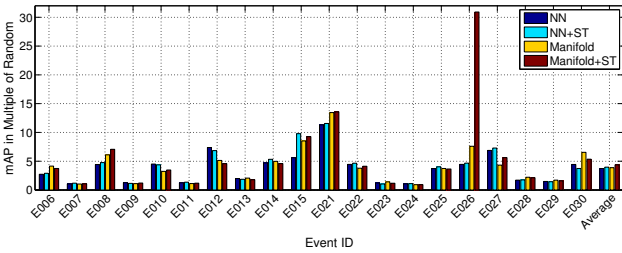
**Baselines:** We compare 5 alternative baselines for TRECVID MED zeroshot event detection.

1. **Random Guess** - Randomly rank the candidates.
2. **NN (X-RR-X-NN)**. Rank videos with  $l_2$  distance in the semantic space.
3. **NN + ST (X-RR-✓-NN)**. Adjust prototypes with self-training.

<sup>3</sup> This is different to some previous studies that trained on the Research Set [Habibian et al \(2014b,a\)](#).

**Table 3** Improving zero-shot action recognition performance (average % accuracy  $\pm$  standard deviation for HMDB51, UCF101 and USAA and mean average precision  $\pm$  standard deviation for Olympic Sports and CCV). \* Our simplified implementation. \*\* As reported by Fu et al (2014a), not quantitatively comparable due to different features.

Our Method				HMDB51	UCF101	Olympic Sports	CCV	USAA
Data Aug	Embed	ST	Match					
X	RR	X	NN	14.5 $\pm$ 2.7	11.7 $\pm$ 1.7	35.7 $\pm$ 8.8	20.7 $\pm$ 3.0	29.5 $\pm$ 5.5
X	RR	✓	NN	17.0 $\pm$ 3.1	15.9 $\pm$ 2.3	37.3 $\pm$ 9.1	21.7 $\pm$ 3.2	30.2 $\pm$ 5.2
X	MR	X	NN	15.9 $\pm$ 3.1	12.9 $\pm$ 2.2	37.7 $\pm$ 9.5	21.4 $\pm$ 3.0	29.8 $\pm$ 4.0
X	MR	✓	NN	18.6 $\pm$ 3.9	17.6 $\pm$ 2.7	<b>38.6<math>\pm</math>10.6</b>	<b>22.5<math>\pm</math>3.4</b>	<b>35.5<math>\pm</math>4.0</b>
X	RR	X	GC	15.3 $\pm$ 2.7	13.5 $\pm$ 1.8	-	-	26.1 $\pm$ 6.7
X	RR	✓	GC	17.0 $\pm$ 2.9	14.8 $\pm$ 2.0	-	-	29.0 $\pm$ 4.0
X	RR	X	NRM	16.1 $\pm$ 2.7	13.9 $\pm$ 1.5	-	-	28.6 $\pm$ 7.2
X	RR	✓	NRM	17.2 $\pm$ 2.9	16.1 $\pm$ 2.2	-	-	28.6 $\pm$ 7.6
X	MR	X	NRM	18.0 $\pm$ 3.2	15.6 $\pm$ 2.0	-	-	28.2 $\pm$ 5.4
X	MR	✓	NRM	<b>19.1<math>\pm</math>3.8</b>	<b>18.0<math>\pm</math>2.7</b>	-	-	31.6 $\pm$ 3.2
✓	MR	X	NN	19.4 $\pm$ 3.0	17.3 $\pm$ 1.8	47.7 $\pm$ 8.4	33.2 $\pm$ 4.0	28.2 $\pm$ 4.6
✓	MR	✓	NN	23.5 $\pm$ 3.7	22.8 $\pm$ 2.5	<b>51.1<math>\pm</math>8.7</b>	<b>35.5<math>\pm</math>4.6</b>	<b>42.8<math>\pm</math>8.7</b>
✓	MR	X	NRM	21.3 $\pm$ 2.6	20.1 $\pm$ 1.8	-	-	35.6 $\pm$ 2.6
✓	MR	✓	NRM	<b>23.7<math>\pm</math>3.4</b>	<b>23.9<math>\pm</math>2.7</b>	-	-	42.6 $\pm$ 9.1
Alternatives								
Random Chance				4.0	2.0	12.5	10.0	25.0
DAP Lampert et al (2009, 2014)				-	15.2 $\pm$ 1.9	40.8 $\pm$ 11.4	-	37.9 $\pm$ 5.9
IAP Lampert et al (2009, 2014)				-	<b>15.6<math>\pm</math>2.2</b>	40.8 $\pm$ 11.0	-	31.7 $\pm$ 1.6
HAA* Liu et al (2011)				-	14.3 $\pm$ 2.0	<b>41.9<math>\pm</math>11.7</b>	-	<b>41.2<math>\pm</math>9.8</b>
M2LATM** Fu et al (2014a)				-	-	-	-	41.9



**Fig. 3** Zeroshot performance on TRECVID MED 2013. Numbers are multiples of random mAP for each event.

4. **Manifold** (X-MR-X-NN). Add manifold regularization term in the visual to semantic regression model.
5. **Manifold + ST** (X-MR-✓-NN) - manifold regularization regression with self-training.

We were not able to investigate data augmentation for TRECVID due to the different feature encoding from the other action datasets.

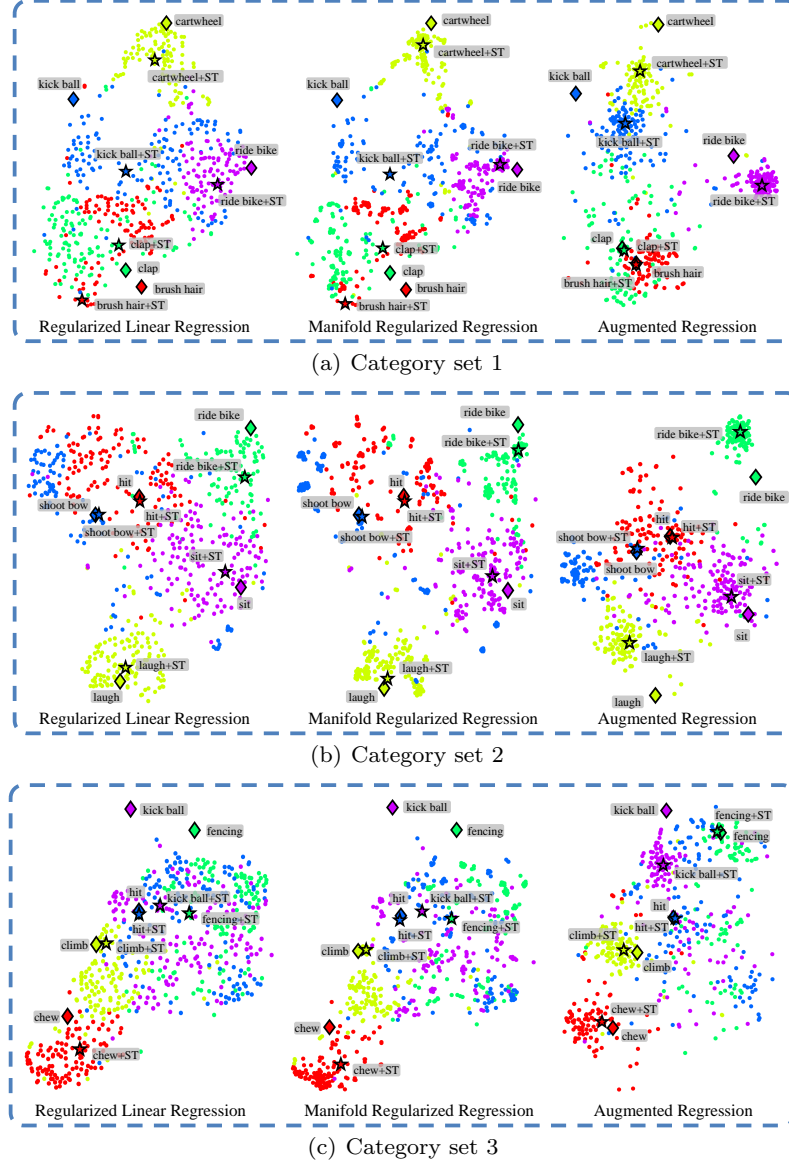
We present the performance of zeroshot learning on TRECVID MED 2013 in Fig. 3. The performance of 4 alternative models are reported in terms of multiple of random guess performance (since the base retrieval precision depends strongly on the overall prevalence). Compared to Random Guess (0.28%), our direct embedding approach (NN) is very effective at zero-shot video detection. Self-Training and manifold regularization further improve the performance.

#### 4.4 Zero-Shot Qualitative Visualization

In this section we illustrate qualitatively the effect of our contributions on the resulting embedding space matching problem. For visualisation, we randomly sample 5 testing classes from HMDB51 and project all samples from these classes into the semantic space by (i) conventional ridge regression; (ii) manifold regularized regression and (iii) manifold regularised ridge regression with data augmentation. The results are visualised in 2D with t-SNE Maaten and Hinton (2008). Three sets of testing classes are presented for diversity. Data instances are shown as dots, prototypes (class name projections) as diamonds, and self-training adapted prototypes as stars. Colours indicate category.

There are three main observations from Fig. 4: (i) Manifold regularised regression yields better visual-semantic projections as instances of the same class tend to form tighter clusters. This is due to the constraint of preserving the manifold structure from the visual feature space. (ii) Data augmentation yields an even more accurate projection of unseen data, as instances are projected closer to the prototypes and classes are more separable. (iii) Self-training is effective as the adapted prototypes (stars) are closer to the center of the corresponding samples (dots) than the original prototypes (diamonds). These observations illustrate the mecha-





**Fig. 4** A qualitative t-SNE illustration of ZSL with semantic space representation for random testing class subsets (a), (b) and (c). Variants: ridge regression, manifold regression and data augmented manifold regression. Dots indicate instances, color categories, and star/diamond show category prototypes with/without self-training.

nism of our ZSL accuracy improvement on conventional approaches.

#### 4.5 Understanding ZSL and Predicting Transferrability

In this section we present further insight into considerations on what factors will affect the efficacy of ZSL, through a category-level analysis. The basic assumption of ZSL is that the embedding  $f(x)$  trained on known class data, will also apply to testing classes. As we have discussed throughout this study, this assumption is stretched to some extent due to the disjoint training and testing category sets. This leads us to investi-

gate how zero-shot performance depends on the specific choice of training classes and their relation to the held out testing classes.

#### Impact of training class choice on testing performance:

We first investigate whether there are specific classes which, if included as training data, significantly impact testing class performance. To study this, we compute the correlation between training class inclusion and testing performance. Specifically consider a pair of random variables  $\{y_i^{tr}, y_j^{te}\}$  where  $y_i^{tr}$  is a binary scalar indicating if the  $i$ th class is in the training set and  $y_j^{te}$  is the recognition accuracy of the  $j$ th testing class. We compute the correlation  $corr(i, j)$  between

every pair of variables over the 50 random splits:

$$\text{corr}(i, j) = \frac{\mathbb{E}[(y_i^{tr} - \overline{y_i^{tr}})(y_j^{te} - \overline{y_j^{te}})]}{\text{var}(y_i^{tr})\text{var}(y_j^{te})}. \quad (16)$$

We use chord diagrams to visualize the relation between categories in Fig 5(a). The strength of positive cross-category correlation is indicated by the width of the bands connecting the categories on the circle. I.e., a wide band indicates inclusion of one category as training data facilitates the zero-shot recognition of the other<sup>4</sup>.

We observe from the Chord Diagram that some category pairs interact strongly. For example  $\{\text{climb stairs-climb}\}$ ,  $\{\text{ride horse-ride bike}\}$ ,  $\{\text{situp-pushup}\}$  are classes which mutually support each other. Other categories are supportive only in one direction. For example *Cartwheel* helps *handstand* (but not vice-versa).

**Cross-class transferability correlates with word-vector similarity:** We next investigate the affinity between category names’ vector representations, and whether it is correlated with class transferability. Category name affinities are shown in Fig 5(b), which reflect some of the same transferability interactions between classes. To quantify the connection between transfer efficacy and classname relatedness, we compute the correlation coefficients between the class-accuracy correlation matrix (Fig 5(a)) and class name-affinity matrix (Fig 5(b)). This is 0.548, suggesting that class name relatedness and efficacy for ZSL are indeed connected. Intuitively, the presence of a given training class provides data in a particular region of the mapping space  $f(x)$  that helps the mapping to perform better in that area, and thus supports *related* (nearby) testing categories.

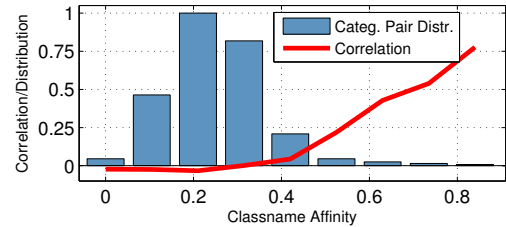
To illustrate this connection, we list the top 10 positively correlated category pairs in Table 6. Here the correlation of action 1 being in training and action 2 in testing is given as *Fwd Corr*, with *Back Corr* being the opposite. The affinity between category names are given as *WV Aff* which is defined as percentile rank of word-vector distance (closer to 1 means more similar). Clearly highly correlated categories have higher word-vector similarity.

Although zero-shot transfer overall is effective, there are also some individual negative correlations. We illustrate the distribution of positive and negative transfer outcomes in Fig. 6. Here we sort all the class pairings into ten bins by their name affinity and plot the resulting histogram (blue bars). Clearly the majority of

**Table 6** Top 10 positive correlated class pairs

Action 1	Action 2	Fwd Corr	Back Corr	WV Aff
climb stairs	climb	0.94	0.92	0.98
ride horse	ride bike	0.95	0.91	0.98
situp	pushup	0.96	0.79	0.91
sword exercise	sword	0.87	0.85	0.98
handstand	cartwheel	0.62	0.96	0.97
eat	drink	0.75	0.81	0.96
smile	laugh	0.82	0.72	0.97
walk	run	0.61	0.90	0.96
shoot ball	dribble	0.52	0.87	0.97
sword	draw sword	0.86	0.45	0.98

pairs have low classname affinity. For each bin of class-pairs, we also compute their average correlation defined in Eq (13) (Fig. 6, red line). There are a few observations to be made: (i) Class name affinity is clearly related to positive correlation: the correlation (red line) goes up significantly for high-affinity class pairs. (ii) There are a relatively small number of category pairs that account for the high positive correlation outcomes (low blue bars to the right). This suggests that overall ZSL efficacy is strongly impacted by the presence of key supporting classes in the training set. (iii) There are a larger number of category pairs which exhibit negative transferability (red correlation is negative around affinity of 0.2). However negative transfer effects are quantitatively weak compared to positive transfer (red correlation line gets only weakly negative but strongly positive).



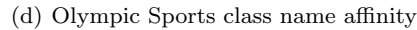
**Fig. 6** The connection between transfer efficacy and class-name affinity: Illustrated by class correlation v.s. class name affinity.

**Predicting Transferability:** Based on the previous observations we hypothesize that class name affinity is predictive of ZSL performance, and may provide a guide to selecting an appropriate set of training classes to maximise ZSL efficacy. Define the relatedness of a putative training class  $y_i$  to the set of testing classes  $\{y_j\}_{j \in T_{\text{test}}}$  as the maximal class name affinity:

$$R(y_i, T_{\text{test}}) = \max_{j \in T_{\text{test}}} (1 - d(g(y_i), g(y_j))) \quad (17)$$

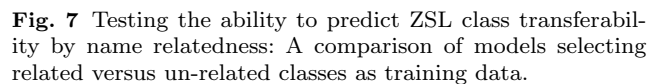
where we use cosine distance for  $d(\cdot)$ . This name-affinity metric provides a means to evaluate any potential train-

<sup>4</sup> Due to the large number of categories we apply two pre-processing steps before plotting: (1) Convert all correlation coefficients to positive value by exponentially power scaling the correlation coefficient; (2) Remove highly negative correlated pairs to avoid clutter.



ing class for its relevance to the testing set. We use the name affinity-based relatedness metric to test the ability to predict transferability outcome and hence construct a good training set for a particular set of testing classes.

els as  $X$  varies, where *Related* selects the top  $X\%$  and *Unrelated* the bottom  $100 - X\%$ .



The main observation are as follows: (i) A crossover happens at 30%, which means the model learned on the 30% subset of related training classes outperforms the model learned on the much larger 70% of unrelated classes. (ii) As more classes are included the *Related* model increases in performance more rapidly than the unrelated one, and saturates after the top 50% are included. Both of these observations demonstrate that the related classes are more valuable than the unrelated ones (as the crossover is to the left of 50%), and that semantic relatedness of the training set is predictive of the efficacy at testing time.

#### 4.6 Multi-Shot Learning

We have thus far focused on the efficacy of unsupervised word-vector embeddings for zero-shot learning. In this section we verify that the same representation also performs comparably to state-of-the-art for standard supervised (multi-shot) action recognition. We use the standard data splits and evaluation metrics for all 4 datasets.

**Alternatives:** We compare our approach to:

1. **Low-Level Feature** Wang and Schmid (2013) the state-of-the-art results based on low-level features.
2. **Human-Labelled Attribute (HLA)** Zheng and Jiang (2014) Exploits an alternative semantic space using human labelled attributes. The model trains binary linear SVM classifiers for attribute detection and uses the vector of attribute scores as a representation. A SVM classifier with RBF kernel is then trained on attribute representation to predict final labels.
3. **Data Driven Attribute (DDA)** Zheng and Jiang (2014) Learns attributes from data using dictionary learning. These attributes are complementary to the human labelled ones. Automatically discovered attributes are processed in the same way as HLA for action recognition.
4. **Mixed attributes (Mix)** Zheng and Jiang (2014) A combination of HLA and DDA is applied to exploit the complementary information in two attribute sets.
5. **Semantic embedding model (Embedding)** first learns a word-vector embedding based on regularised linear regression, as in ZSL. But the standard supervised learning data-split is adopted. All data are mapped into the semantic space via regression and a linear SVM classifier is trained for each category with the mapped training data.

The resulting accuracies are shown in Table 7. We observe that our semantic embedding is comparable to the state-of-the-art low-level feature-based classification and is comparable or slightly better than the conventional attribute-based intermediate representations despite the fact that no supervised manual attribute definition and annotation is required.

#### 4.7 Efficiency and Runtime

The efficiency of our ZSL algorithm compares favourably against existing alternatives due to its closed-form solution to mapping the visual feature space to word-vector semantic embedding space (Eq. 9). For instance, it took about 300 seconds to train and test 50 splits of the entire HMDB51 benchmark dataset (6766 videos of 51 categories of actions), or 520 seconds with data augmentation, using a single thread on a Intel E5-2680 CPU. The computational cost is dominated by the matrix inversion in Eq. 9, which can be sped up by exploiting efficient matrix libraries.

### 5 Detailed Parameter Sensitivity Analysis

In the main experiments we set the free parameters ridge regularizer  $\gamma_A = 1^{-6}$ , manifold regularizer  $\gamma_I = 40$ , manifold Knn graph  $N_K^G = 5$ , Self-Training Knn  $N_K^{st} = 100$ . In this section we analyse the impact of these free parameters in our model.

#### 5.1 word-vector Dimension

We investigate how the specific word-vector model  $z = g(y)$  affects the performance of our framework. For the study of word-vector dimension we train word-vectors on 4.6M Wikipedia documents<sup>5</sup> and vary dimension from 32 to 1024. We then evaluate the performance of zeroshot and multishot learning v.s. different dimension of embedding space. The results are given in Fig. 8.

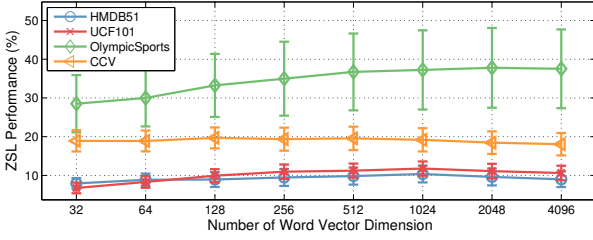
We observe that word-vector dimension does affect the zero-shot recognition performance. Performance generally increases with dimension of word-vector from 32 to 4096 in HMDB51, UCF101 and Olympic Sports, while showing no clear trend for CCV. In general a reasonable word-vector dimension is between 256 to 2048.

<sup>5</sup> Google News Dataset is not publicly accessible. So we use a smaller but public dataset - 4.6M Wikipedia documents.



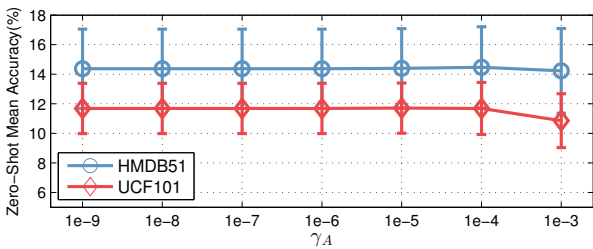
**Table 7** Standard supervised action recognition. Average accuracy for HMDB51 and UCF101 datasets. Mean average precision for Olympic Sports and CCV.

Method	HMDB51	UCF101	Olympic Sports	CCV
Low-Level Feature Wang and Schmid (2013)	58.4	84.6	92.1	68.0
HLA Zheng and Jiang (2014)	-	81.7	-	-
DDA Zheng and Jiang (2014)	-	79.0	-	-
Mix Zheng and Jiang (2014)	-	82.3	-	-
Embedding	56.4	82.0	93.4	51.6

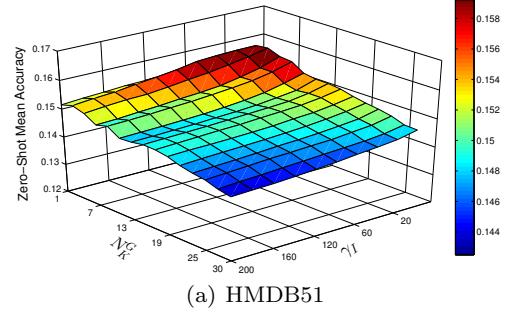
**Fig. 8** Zeroshot performance v.s. dimension of word-vector.

## 5.2 Visual to Semantic Mapping

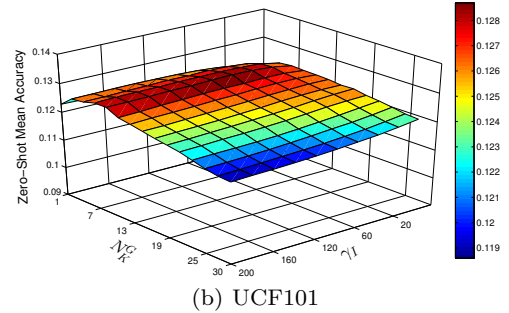
**Ridge regression regularisation:** We learn the visual to semantic mapping with regularized linear regression. The regularization parameter  $\gamma_A$  controls the regression model complexity. Here, we study the impact of  $\gamma_A$  on zero-shot performance. We measure the 50 splits' average accuracy by varying  $\gamma_A$  in the range of  $\{1^{-9}, 1^{-8}, \dots, 1^{-3}\}$ . A plot of zero-shot mean accuracy v.s. regularization parameter is given in Fig. 9. From this figure we observe that our model is insensitive to the ridge parameter.

**Fig. 9** Zero-shot mean accuracy v.s. ridge regression parameter

**Manifold regression:** We have seen that exploiting unlabelled data in manifold learning improves zero-shot performance. Two parameters are involved: the manifold regularization parameter  $\gamma_I$  in Loss function (Eq. 8) and  $N_K^G$  in constructing the symmetrical KNN graph.  $\gamma_I$  controls the preference for preserving the manifold structure in mapping to the semantic space, versus exactly fitting the training data. Parameter  $N_K^G$  determines the precision in modelling the manifold structure.



(a) HMDB51



(b) UCF101

**Fig. 10** Zero-shot recognition accuracy with respect to manifold regression parameters  $\gamma_I$  and  $N_K^G$ .

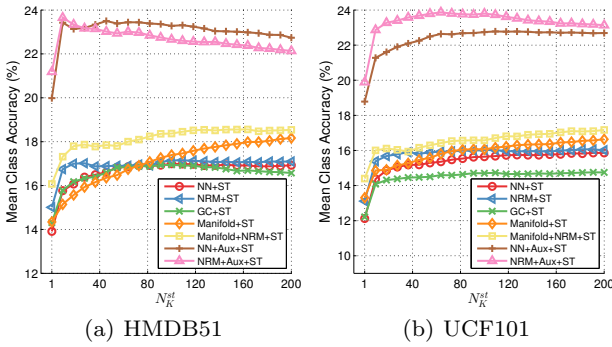
Small  $N_K^G$  may more precisely exploit the testing data manifold, however it is more prone to noise in the neighbours.

Here we analyse the impact of these two parameters,  $\gamma_I$  and  $N_K^G$  by measuring zero-shot recognition accuracy on HMDB51 and UCF101. We evaluate the joint effect of  $\gamma_I$  and  $N_K^G$  while fixing  $\gamma_A = 1^{-6}$ . Specifically we test  $\gamma_I \in \{20, 40, \dots, 100\}$  and  $N_K^G \in \{1, 3, 5, \dots, 29\}$ . The results in Fig. 10 show that there is a slightly preference towards moderately low values of  $N_K^G$  and  $\gamma_I$ , but the framework is not very sensitive to these parameters.

## 5.3 Self-Training

We previously demonstrated in Table 3, that self-training (Section 3.4) helps to mitigate the domain shift problem. Here, we study the influence of the  $N_K^{st}$  parameter for KNN in self-training. Note the  $N_K^{st}$  concerns the neighbouring data distribution around prototypes at testing time rather than manifold regularization KNN graph  $N_K^G$  at training time. We evaluate  $N_K^{st} \in \{1, 2, 3, \dots, 200\}$ .

To thoroughly examine the effectiveness of self-training, we investigate all baselines with self-training including X-RR- $\checkmark$ -NN (NN+ST), X-RR- $\checkmark$ -NRM (NRM+ST), X-RR- $\checkmark$ -GC (GC+ST), X-MR- $\checkmark$ -NN (Manifold+ST), X-MR- $\checkmark$ -NRM (Manifold+NRM+ST),  $\checkmark$ -RR- $\checkmark$ -NN (NN+Aux+ST) and  $\checkmark$ -RR- $\checkmark$ -NRM (NRM+Aux+ST) introduced in section 4.2. The accuracy v.s.  $N_K^{st}$  is illustrated in Fig. 11. Performance is robust to  $N_K^{st}$  when  $N_K^{st}$  is above 20.



**Fig. 11** Zero-shot recognition accuracy v.s. self-training parameter  $K$ .

## 6 Conclusion

In this study, we investigated *unsupervised* word-vector embedding space representation for zero-shot action recognition for the first time. The fundamental challenge of zero-shot learning is the disjoint training and testing classes, and associated domain-shift. We explored the impact of four simple but effective strategies to address this: data augmentation, manifold regularisation, self-training and hubness correction. Overall we demonstrated that these strategies are complementary, and together facilitate a highly effective system that outperforms significantly existing methods for zero-shot recognition despite their use of strongly *supervised* embeddings (attributes). Moreover, our model has a closed-form and is very simple to implement (a few lines of matlab) and very efficient to run compared to existing state-of-the-art ZSL methods. Finally, we also provide a unique analysis of the inter-class affinity for ZSL, giving insight into why and when ZSL works. This provides for the first time two new capabilities: the ability to predict the efficacy of a given ZSL scenario in advance, and a mechanism to guide the construction of suitable training sets for a desired set of target classes.

## References

- Aggarwal J, Ryoo M (2011) Human activity analysis: A review. *ACM Computer Survey*
- Akata Z, Reed S, Walter D, Lee H, Schiele B (2015) Evaluation of output embeddings for fine-grained image classification. In: *CVPR*
- Belkin M, Niyogi P, Sindhvani V (2006) Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *The Journal of Machine Learning Research*
- Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In: *ICCV*
- Dinu G, Lazaridou A, Baroni M (2015) Improving zero-shot learning by mitigating the hubness problem. In: *ICLR, Workshop Track*
- Fu Y, Hospedales TM, Xiang T, Gong S (2012) Attribute learning for understanding unstructured social activity. In: *ECCV 2012*
- Fu Y, Hospedales TM, Xiang T, Gong S (2014a) Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Fu Y, Yang Y, Gong S (2014b) Transductive Multi-label Zero-shot Learning. In: *BMVC*
- Fu Y, Hospedales TM, Xiang T, Gong S (2015a) Transductive Multi-view Zero-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Fu Z, Xiang T, Kodirov E, Gong S (2015b) Zero-shot object recognition by semantic manifold distance. In: *CVPR*
- Gan C, Lin M, Yang Y, Zhuang Y, GHauptmann A (2015) Exploring semantic inter-class relationships (sir) for zero-shot action recognition. In: *AAAI*
- Habibian A, Mensink T, Snoek CG (2014a) Composite concept discovery for zero-shot video event detection. In: *ICMR*
- Habibian A, Mensink T, Snoek CGM (2014b) VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In: *ACM Multimedia*
- Jiang Y, Wu Z, Wang J, Xue X, Chang S (2015) Exploiting feature and class relationships in video categorization with regularized deep neural networks. *arXiv preprint arXiv:150207209*
- Jiang YG, Ye G, Chang SF, Ellis DPW, Loui AC (2011) Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: *ICMR*
- Jiang YG, Liu J, Roshan Zamir A, Laptev I, Piccardi M, Shah M, Sukthankar R (2013) THUMOS challenge: Action recognition with a large number of classes

- Klaser A, Marszałek M, Schmid C (2008) A spatio-temporal descriptor based on 3d-gradients. In: BMVC
- Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: A large video database for human motion recognition. In: ICCV
- Lampert CH, Nickisch H, Harmeling S (2009) Learning to detect unseen object classes by between-class attribute transfer. In: CVPR
- Lampert CH, Nickisch H, Harmeling S (2014) Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Laptev I (2005) On space-time interest points. *International Journal of Computer Vision*
- Larochelle H, Erhan D, Bengio Y (2008) Zero-data learning of new tasks. In: AAAI
- Lazaridou A, Bruni E, Baroni M (2014) Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In: ACL
- Liu J, Kuipers B, Savarese S (2011) Recognizing human actions by attributes. In: CVPR
- Maaten LVD, Hinton G (2008) Visualizing data using t-SNE. *Journal of Machine Learning Research*
- Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: CVPR
- Mensink T, Gavves E, Snoek CG (2014) Costa: Co-occurrence statistics for zero-shot classification. In: CVPR
- Mikolov T, Sutskever I, Chen K (2013) Distributed representations of words and phrases and their compositionality. In: NIPS
- Milajevs D, Kartsaklis D, Sadrzadeh M, Purver M (2014) Evaluating neural word representations in tensor-based compositional settings. In: EMNLP
- Mitchell J, Lapata M (2008) Vector-based Models of Semantic Composition. *Computational Linguistics*
- Niebles CWWFL Juan Carlos Chen (2010) Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV
- Over P, Fiscus J, Sanders G, Joy D, Michel M, Smeaton-Alan AF, Quénot-Georges G (2014) Trecvid 2013—an overview of the goals, tasks, data, evaluation mechanisms, and metrics
- Palatucci M, Hinton G, Pomerleau D, Mitchell TM (2009) Zero-shot learning with semantic output codes. In: NIPS
- Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*
- Perronnin F, Sánchez J, Mensink T (2010) Improving the fisher kernel for large-scale image classification. In: ECCV
- Poppe R (2010) A survey on vision-based human action recognition. *Image and vision computing*
- Romera-Paredes B, Torr PHS (2015) An embarrassingly simple approach to zero-shot learning. In: ICML
- Scholkopf B, Smola AJ (2002) *Learning with kernels*. MIT Press. (2002),
- Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: A local svm approach. In: ICPR
- Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: ACM Multimedia
- Shao L, Zhu F, Li X (2015) Transfer learning for visual categorization: a survey. *IEEE transactions on neural networks and learning systems*
- Socher R, Ganjoo M (2013) Zero-shot learning through cross-modal transfer. In: NIPS
- Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:12120402
- Wang H, Schmid C (2013) Action Recognition with Improved Trajectories. In: ICCV
- Wang H, Kläser A, Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*
- Wang H, Oneata D, Verbeek J, Schmid C (2015) A robust and efficient video representation for action recognition. *International Journal of Computer Vision*
- Xu X, Hospedales Hospedales T, Gong S (2015) Semantic embedding space for zero shot action recognition. In: ICIP
- Yang Y, Hospedales T (2015) A unified perspective on multi-domain and multi-task learning. In: ICLR
- Yeffet L, Wolf L (2009) Local trinary patterns for human action recognition. In: ICCV
- Zhao F, Huang Y, Wang L, Tan T (2013) Relevance topic model for unstructured social group activity recognition. In: NIPS
- Zheng J, Jiang Z (2014) Submodular Attribute Selection for Action Recognition in Video. In: NIPS